Validation of Imaging Biomarkers for Response Evaluation in Lung and Prostate Cancer

Kramer, G.M.

2019

document version
Publisher's PDF, also known as Version of record

CHAPTER 4

Repeatability of Quantitative Whole Body ^{18}F-FDG PET/CT Uptake Measures as Function of Uptake Interval and Lesion Selection in Non-Small Cell Lung Cancer Patients

G.M. Kramer*, V. Frings*, N. Hoetjes, O.S. Hoekstra, E.F. Smit,
A.J. de Langen, R. Boellaard

** G.M. Kramer and V. Frings contributed equally to this work.*

J Nucl Med. 2016;57(9):1343-1349

ABSTRACT

Objectives: Change in ^{18}F-FDG uptake may predict response to anticancer treatment. The PET Response Criteria in Solid Tumors (PERCIST) suggest a threshold of 30% change in standardized uptake value (SUV) to define partial response and progressive disease. Evidence underlying these thresholds consists of mixed stand-alone PET and PET/CT data with variable uptake intervals and no consensus on the number of lesions to be assessed. Additionally, there is increasing interest in alternative ^{18}F-FDG uptake measures such as metabolically active tumor volume (MATV) and total lesion glycolysis (TLG). The aim of this study was to comprehensively investigate the repeatability of various quantitative whole body ^{18}F-FDG metrics in non-small cell lung cancer (NSCLC) patients as a function of tracer uptake interval and lesion selection strategies.

Methods: Eleven NSCLC patients, with at least one intrathoracic lesion ≥ 3cm, underwent double baseline whole body ^{18}F-FDG PET/CT scans at 60 and 90 minutes post-injection (p.i.) within 3 days. All ^{18}F-FDG avid tumors were delineated with an 50% threshold of SUV_{peak} adapted for local background. SUV_{max}, SUV_{mean}, SUV_{peak}, TLG, MATV, and tumor-to-blood and -liver ratios were evaluated, as well as the influence of lesion selection and two methods for correction of uptake time differences.

Results: The best repeatability was found using the SUV metrics of the averaged PERCIST target lesions (repeatability coefficients < 10%). The correlation between test and retest scans was strong for all uptake measures at either uptake interval (ICC > 0.97 and R^2 > 0.98). There were no significant differences in repeatability between data obtained 60 and 90 minutes p.i.. When only PERCIST defined target lesions were included (n = 34) repeatability improved for all uptake values. Normalisation to liver or blood uptake or glucose correction did not improve repeatability. However, after correction for uptake time the correlation of SUV measures and TLG between the 60 and 90 minutes data significantly improved without affecting test-retest performance.

Conclusion: This study suggests that a 15% change of SUV_{mean}/SUV_{peak} at 60 minutes p.i. can be used to assess response in advanced NSCLC patients if up to 5 PERCIST target lesions are assessed. Lower thresholds could be used in averaged PERCIST target lesions (< 10%).

INTRODUCTION

^{18}F-fluorodeoxyglucose (^{18}F-FDG) is widely used as a diagnostic or prognostic tool in oncology, but its role as biomarker of response to cancer therapy is less well established (*1-5*). Evaluation of response using positrion emmison tomography (PET) can be performed visually (e.g. International Conference on Malignant Lymphoma taxonomy in malignant lymphoma (*6*)) or (semi)quantitatively (*2*). For the latter, the proposed PET Response Criteria in Solid Tumors (PERCIST) suggest a threshold of 30% change in standardized uptake value (SUV) (combined with a minimal absolute change) to define either partial response or progressive disease (*3*). Evidence underlying these thresholds consists of mixed test-retest data from stand-alone PET and PET/CT scanners, with variable uptake intervals (*3,7*). To date, optimal tracer uptake time for response assessment is still matter of debate (60 vs. 90 min post-injection). Furthermore, traditionally most repeatability studies reported on solitary tumor measurements (*7*). Yet, thresholds for response evaluation should also apply to patients with multiple metastases. The PERCIST suggest measuring up to 5 lesions for response assessment; however the impact of lesion selection strategies on repeatability requires further research.

More recently, two PET/CT studies performed in ovarian and non-small cell lung cancer (NSCLC) patients showed conflicting results on repeatability of SUV measurements (*8,9*). As discussed, only SUV metrics were assessed in these studies and the effect of normalization to blood or liver SUV has not been evaluated. This, however, could improve repeatability of the uptake metrics (*10,11*). Finally, there is an increasing interest in alternative ^{18}F-FDG uptake measures, such as total lesion glycolysis (TLG) and metabolically active tumor volume (MATV) (*3,12*). There are only limited data on the test-retest performance of these uptake metrics and influence of uptake time interval and lesion selection have not been investigated to our knowledge.

The aim of this study was therefore to comprehensively investigate the repeatability of various quantitative whole-body ^{18}F-FDG uptake and volumetric measures in advanced NSCLC patients as a function of tracer uptake interval and lesion selection strategy. Furthermore, we evaluated two proposed methods to account for variable uptake intervals.

MATERIALS AND METHODS

Patients

11 NSCLC patients (7 men) with at least one intra-thoracic lesion ≥ 3cm in the largest diameter, who had not received chemotherapy in the past 4 weeks and without known diabetes mellitus, were included between January 2013 and January 2015 by their pulmonary physician in the VU University Medical Center, Amsterdam, The Netherlands. Patients underwent double baseline whole body ^{18}F-FDG PET/CT scans at 60 and 90 minutes post-injection (p.i.). In total, 11 and 10 test-retest scans were obtained at 60 and 90 minutes p.i. respectively (one patient did not undergo one 90 minutes scan due to back pain). There were no significant differences in patient preparation and PET acquisition between the test and retest scans (Table 4.1). This study was approved by the institutional review board and was registered in the Dutch trial register (trialregister.nl, NTR3508). Written informed consent for all subjects was obtained prior to study enrolment.

PET Imaging

All PET scans were obtained using a Philips Gemini TF PET/CT scanner (Philips Healthcare, Eindhoven, The Netherlands). Scans were obtained and reconstructed following the guideline recommendations of the European Association of Nuclear Medicine (*13*). Patients were asked to fast at least 6 hours before the PET scan, and blood glucose levels were measured twice before tracer injection to correct for measurement errors. Patients underwent a low-dose CT during tidal breathing for attenuation correction, followed by a whole body ^{18}F-FDG PET/CT scan (skull vertex to mid-thigh) 60 min p.i. at 2 minutes per bed position. Ninety minutes p.i. a second whole body PET scan was acquired, followed by a second low-dose CT for attenuation correction. This procedure was repeated within 3 days after the first scan. Weight, height, total injected activity, time of injection, residual activity and exact scan start time of both time points were recorded for each session.

Table 4.1: Descriptive statistics of study population, median (range), *p*-values from Wilcoxon signed-rank test.

	Median (range)	**Scan 1**	**Scan 2**	***p*-value**
Age (years)	61 (45-66)			
Patients	11			
Tumortype				
- Adenocarcinoma	8			
- Squamous-cell carcinoma	3			
Tumor stage				
- IIIb	4			
- IV	7			
Lesions (n)				
- Intrathoracal	41			
- Extrathoracal	19			
Time between scans (days)	1 (1-2)			
Length (cm)	173 (162-197)			
Weight (kg)		75 (57-110)	74 (57-114)	0.60
Glucose (mmol/L)		5.6 (4.9-6.5)	5.9 (4.5-7.1)	0.62
Uptake time target 60 min		60 (60-67)	60 (60-63)	0.35
Uptake time target 90 min		93 (90-97)	90 (90-95)	0.15
Injected activity (MBq)		252 (194-377)	238 (192-329)	0.86
Residual dose (MBq)		0.16 (0.07-1.7)	0.3 (0.02-3.22)	0.33

Data Analysis

Volumes of interest (VOIs) were generated by delineating ^{18}F-FDG avid tumors using a 50% threshold of SUVpeak adapted for local background (in-house developed software). Details on this method were published previously (*12*). Tumors were selected by a nuclear physician. For each VOI, SUV_{max} (maximum SUV), SUV_{mean} (mean SUV), SUV_{peak} (1.2cm^3 spheric region positioned to maximize its mean value), MATV (50% threshold of SUV_{peak} corrected for local background), TLG (product of SUV_{mean} and

MATV), and tumor-to-blood and tumor-to-liver ratios were determined. The SUV_{mean} of a VOI placed in the ascending aorta (3.3 mL) and liver (14 mL) were used for normalization to blood and liver uptake. SUVs were corrected for lean body mass using James formula (*13*) and all uptake measures were assessed with and without glucose correction.

We applied two methods for correction of uptake metrics for uptake time differences as described by van den Hoff et al. (*14*). The first corrects the 90 minutes data to 60 minutes by estimating the 60 minutes SUVs using

$$SUV_0 = SUV_T \times \left[\frac{SUR_0}{SUR_T} \times \left(\frac{T_0}{T} \right)^{-b} \right] \quad \text{Eq (2)}$$

With

$$SUR_0 = \frac{T_0}{T}(SUR_T - \overline{V}_R) + \overline{V}_R$$

Here SUR represents the tumor-to-blood uptake value and V_R the apparent volume of distribution, which was set to 0.53 and the time exponent *b* was set to 0.313 according to van den Hoff et al. We also determined exponent *b* for our study population using the group-averaged blood activity resulting in a *b* value of 0.5. The second method to correct SUV for uptake time is based on the rule of thumb that $SUR_0/SUR_T \approx (T_0/T)$ resulting in:

$$SUV_0 = SUV_T \times \left(\frac{T_0}{T} \right)^{1-b} \quad \text{Eq (3)}$$

Statistical Analysis

We determined repeatability by calculating the mean and standard deviation (SD) of the absolute and percentage differences between the test and retest scan. Percentage difference was calculated as:

$$\%difference = \frac{scan2 - scan1}{(scan1 + scan2)/2} * 100 \quad \text{Eq (4)}$$

The reproducibility coefficient (RC) was calculated as 1.96 x SD of the percentage and absolute differences for all uptake metrics at both time points. Normality was assessed using a quantile-quantile plot and histogram analyses. A paired *t* test was used to test for significant differences in mean uptake between the test and retest scan and the Levene's test was performed to investigate whether differences in RC were significant. Additionally, linear regression analyses, intraclass correlation coefficient (ICC), and Bland-Altman plots were used to evaluate repeatability.

Repeatability of SUV metrics, MATV, and TLG were evaluated as function of uptake interval, glucose correction and normalisation procedures for SUV metrics (tumor-to-blood and tumor-to-liver ratios). Various lesion selection strategies were applied and their effect on repeatability was evaluated: all lesions, lung tumors, the lesion with the highest uptake per scan, lesions > 4.2mL (diameter: >2cm), PERCIST target lesions (*3*) and averaged PERCIST target lesions. PERCIST target lesions are the 5 hottest lesions with a maximum of 2 per organ and a SUV_{max} > 1.5 x mean liver SUV + 2 SDs per patient. The uptake values of individual PERCIST target lesions were averaged per patient to obtain the averaged PERCIST target lesions. Finally, the effect of correction for uptake time was assessed. Statistical analyses were performed using SPSS (SPSS, Chicago, IL, USA).

RESULTS

Repeatability of Uptake Metrics

Test-retest variability was analysed in 9 NSCLC patients (stage IV) with a total of 60 lesions (Table 4.1). Two patients were excluded from the analysis, one because no retest scan at 90 minutes was obtained and the second due to movement during the retest scan at 60 minutes (mean difference for all uptake measures > 2 SDs). Data including the latter patient are shown in the supplemental section.

Test SUV, TLG and MATV values were plotted against their equivalent retest counterparts. Correlations between the test and retest scans were strong for all uptake measures (ICC > 0.98; CI: 0.97-1.00 and R^2 > 0.97) (Figure 4.1; Table 4.2). SUV_{mean} and SUV_{peak} showed best test-retest performance. Variability of these SUV measures were

not significantly different for the 60 and 90 minutes datasets (RCs 19.9%–26.6% and 15.8%-23.3% respectively) (Figure 4.2; Table 4.3). RCs of the absolute differences ranged from 0.8 for SUV_{mean} to 1.6 for SUV_{max} in the 60 minutes data and from 0.9 to 2.1 in the 90 minutes data. Furthermore, Bland Altman plots showed a correlation between the relative variability and SUV (Figure 4.1), however no correlation was found for the absolute RC.

TLG and MATV showed higher test-retest variability than SUV metrics. In the 60 and 90 minutes groups, absolute RCs of MATV were 10.8 and 8.6, respectively. Even though MATV obtained from VOIs, based on a relative threshold of SUV_{peak}, might depend on SUV measures itself, MATV proved not to be correlated to SUV. When repeatability of TLG was assessed, absolute RCs of 62.4 and 38.2 were found in the 60 and 90 minutes data. Note that these results cannot be directly compared to SUV measures, considering that TLG and MATV have higher values. The absolute and relative differences between the 60 and 90 minutes data were also plotted in a Bland-Altman plot and showed no skewing for any of the uptake metrics.

Table 4.2: Descriptive statistics of the uptake measures for several tissues (mean ± SD).

		PET/CT 60 minutes post-injection (mean ± SD)		PET/CT 90 minutes post-injection (mean ± SD)	
		Test	Retest	Test	Retest
Aorta					
	SUV_{mean}	1.29 ± 0.20	1.24 ± 0.15	1.10 ± 0.17	1.03 ± 0.19
Liver					
	SUV_{mean}	1.69 ± 0.28	1.69 ± 0.24	1.64 ± 0.28	1.60 ± 0.26
Tumor					
	SUV_{max}	8.52 ± 5.40	8.54 ± 5.50	9.87 ± 6.43	9.88 ± 6.54
	SUV_{peak}	6.09 ± 4.53	6.11 ± 4.53	6.96 ± 5.29	7.00 ± 5.38
	SUV_{mean}	5.03 ± 3.29	5.08 ± 3.36	5.76 ± 3.91	5.77 ± 4.00
	TLG	175.29 ± 628.50	179.29 ± 642.73	202.51 ± 728.88	202.84 ± 731.88
	MATV	21.12 ± 56.94	20.92 ± 54.77	21.00 ± 56.40	20.60 ± 54.21

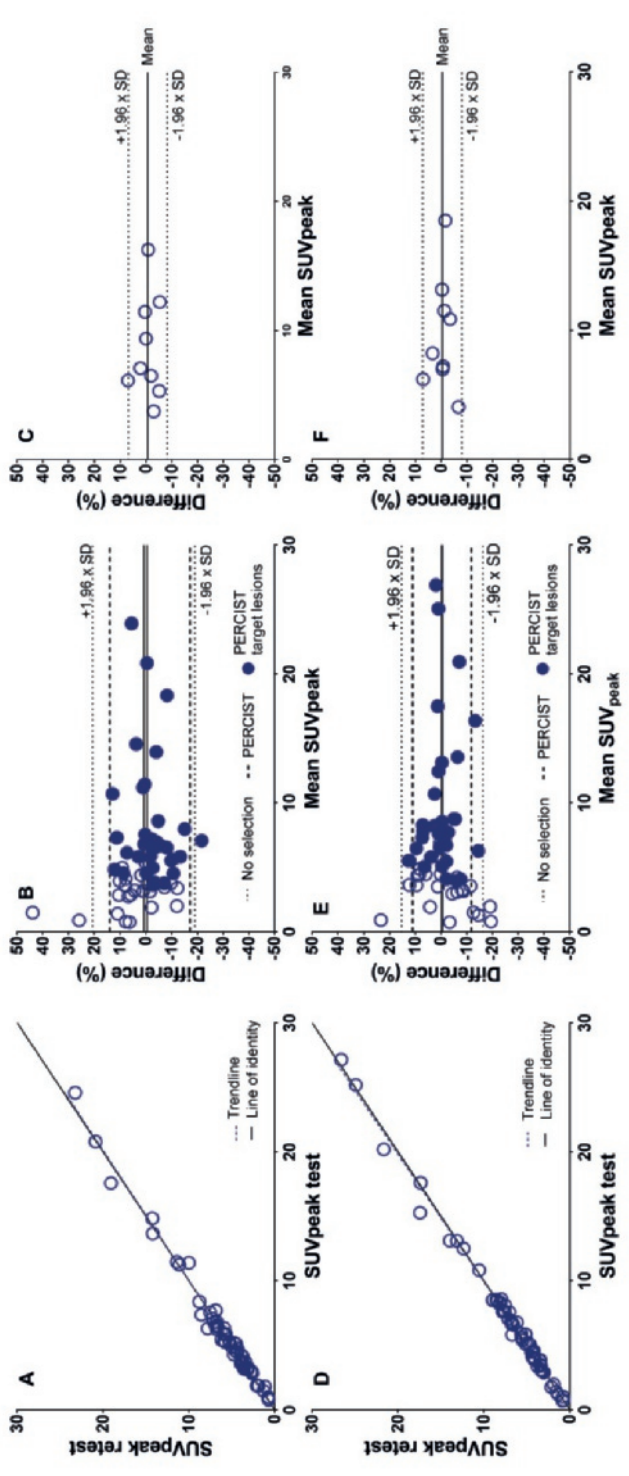


Figure 4.1: SUV_{peak} values of the test scan plotted against those of the retest scan (A and D), the corresponding Bland-Altman plots for the all lesions (B and E) and for the averaged PERCIST target lesions per patient (C and F) are shown. The upper and lower plots represent the 60 minutes data (A-C) and 90 minutes data (D-F) respectively. Plots B and E also show the influence of PERCIST lesion selection criteria on repeatability. (PERCIST target lesions: the 5 hottest lesions per patient; max. 2 per organ; SUV_{max} >1.5 x mean liver SUV +2.

Lesion Selection

Including only PERCIST target lesions (n = 34) improved repeatability both for the 60 and 90 minutes scans (range: 13.8%-15.8% and 11.4%-16.0% respectively) as compared to inclusion of all ^{18}F-FDG avid lesions. These results further improved using the average SUV_{max} SUV_{mean} and SUV_{peak} value of the PERCIST target lesions within 1 patient (Figure 4.2; Table 4.3). When we considered only PERCIST target lesions, RCs of the absolute differences slightly increased (< 0.3) but did not exceed 2.3. For averaged PERCIST target lesions RCs decreased and ranged from 0.8 to 1.3 for the 60 minutes and from 0.4 to 1.2 for the 90 minutes data. Repeatability remained worse in MATV and TLG when only PERCIST target lesions were evaluated. In the 60 and 90 minutes data, RCs of MATV for PERCIST target lesions equalled 14.2 and 11.3 mL. For TLG we found RCs of 82.9 and 50.8 for the 60 and 90 minutes data respectively.

If only lesions showing the highest uptake were included in the analysis, results equalled the averaged PERCIST data and were not influenced by the outlier. Test-retest variability obtained for lung lesions and lesions > 4.2 mL was similar to those of PERCIST target lesions. Moreover, both intra- and extrathoracic lesions were included and no differences in repeatability were found depending on tumor location.

Normalisation to Blood or Liver Uptake and Glucose Correction

Hepatic ^{18}F-FDG uptake was independent of uptake interval and showed low inter-scan variability between the test and retest scans (median: 0.01; IQR: 0.09). Normalisation of SUV to liver uptake did not affect repeatability for any of the uptake measures and times. Normalisation to blood uptake did not influence repeatability for the 60 minutes uptake time, but in the 90 minutes data RCs increased from 17.1% ± 4.0% to 29.6% ± 3.0%. Furthermore, plasma glucose correction adversely affected repeatability at both time points.

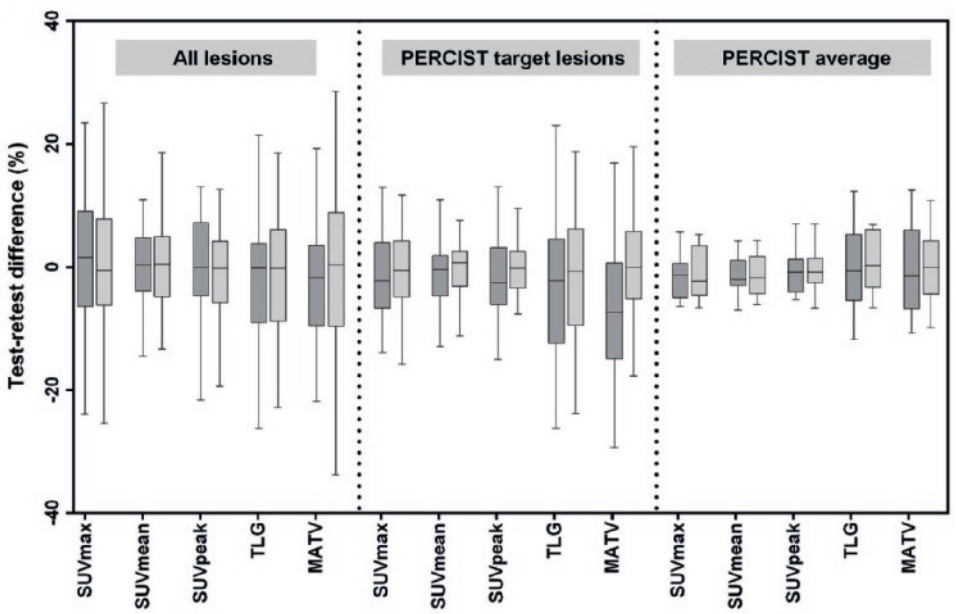


Figure 4.2: Boxplots of the percentage differences between the test and retest scans as obtained from the 60 (dark) and 90 (light) minutes data. The effect of lesion selection and averaging on different ^{18}F-FDG uptake metrics is shown.

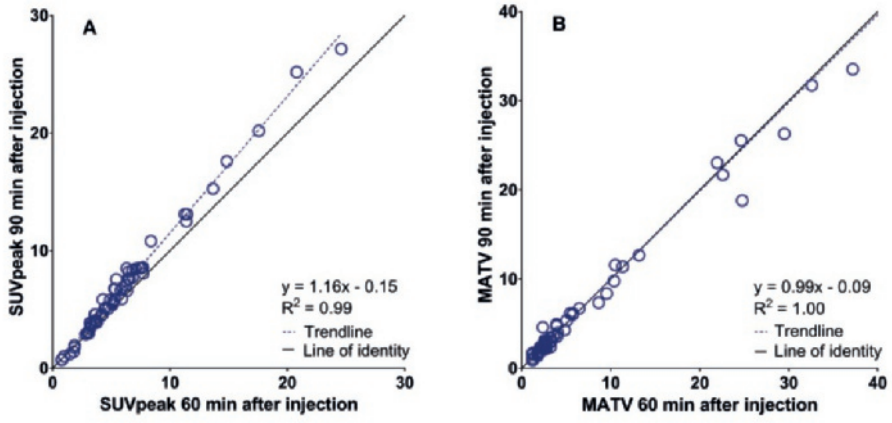


Figure 4.3: SUV_{peak} (A) and MATV (B) values of data obtained 60 minutes post-injection plotted against data obtained 90 minutes post-injection of the test-scan; MATV: Metabolically active tumor volume.

Table 4.3: The mean relative differences and repeatability coefficients (RC) for several uptake metrics and the influence of different uptake times and lesions selection.

		All lesions		*PERCIST-criteria ł*		*PERCIST-criteria averaged*	
Uptake time	*Quantitative tracer uptake measures*	*Mean diffference (%)*	*RC (%)*	*Mean diffference (%)*	*RC (%)*	*Mean diffference (%)*	*RC (%)*
60 minutes post-injection	SUV_{max}	2.2	26.6	-1.8	13.8	-1.5	7.3
	SUV_{mean}	0.7	18.1	-1.7	13.7	-1.3	6.6
	SUV_{peak}	0.8	19.9	-1.6	15.6	-0.7	7.5
	TLG	-1.5	29.0	-1.6	19.1	-0.1	14.7
	MATV	-2.2	30.9	0.0	21.4	-0.2	14.9
90 minutes post-injection	SUV_{max}	-0.6	23.3	-1.1	16.0	-1.2	8.4
	SUV_{mean}	-0.3	17.8	-1.0	11.9	-1.3	6.9
	SUV_{peak}	-0.6	15.8	-0.4	11.4	-0.5	7.6
	TLG	-1.1	23.7	-0.3	15.1	0.9	9.6
	MATV	-0.8	30.7	0.6	20.7	0.1	12.9

ł PERCIST-criteria: 5 hottest lesions, max. 2 per organ and minimum SUV >1.5 x mean liver SUV + 2 SDs

Uptake Time Correction

With the exception of MATV, longer uptake intervals were associated with higher uptake values for both the test and retest data (mean difference range: 8.2- 15.0%) (Figure 4.3). Application of the van den Hoff et al. uptake interval correction method significantly decreased mean differences between the test 60 minutes and retest 90 minutes data for all uptake measures. The 90 minutes data estimated to 60 minutes using equation 2 correlated better with the 60 minutes data than those using the rule-of-thumb (eq. 3). However, mean differences remained more than 5% and were significant. After the *b* value was adjusted to 0.5, correlation further improved and the estimated values no longer differed from the 60 minutes SUV_{mean} and SUV_{peak} data regardless of the lesions included (Figure 4.4). RCs of the percentage difference between the 60 minutes data and the 90 minutes data corrected to 60 minutes were similar to those of corresponding uptake metrics described above.

DISCUSSION

In this study, repeatability of SUV metrics was superior to TLG and MATV and after PERCIST lesion selection criteria were applied RCs improved to < 15%. We observed similar repeatability performance characteristics of several quantitative ^{18}F-FDG uptake measures at 60 and 90 minutes post-injection. The repeatability of ^{18}F-FDG PET has previously been studied, but evidence underlying proposed thresholds for response evaluation consists of mixed test-retest data from stand-alone PET and PET/CT scanners, with variable uptake intervals. Compared with the 25% and 30% thresholds suggested by the European Organization for Research and Treatment of Cancer and PERCIST respectively (*3,15*), we found an improved repeatability for all SUV metrics. Our data are consistent with a study on ^{18}F-FDG PET/CT in patients with recurrent ovarian carcinoma reporting RCs of 16.3% and 17.3% for SUV_{mean} and SUV_{max} (*8*). In addition, similar results were observed in a meta-analysis of mainly intrathoracic lesions (*7*).

A recent multi-center study evaluated ^{18}F-FDG PET/CT in 74 NSCLC patients accrued at 24 different sites (*9*). In contrast to the former studies, here a threshold of 28% decrease and 39% increase for SUV_{max} (32% decrease and 47% increase for SUV_{peak}) were found to reflect true therapeutic effects if per patient only one lesion > 2 cm, with the highest SUV_{max} (> 4 g/mL) was included. These results are comparable to those shown in a multi-center study performed by Velasquez et al. (*16*), suggesting that repeatability of ^{18}F-FDG PET might be more limited in a multi-center setting. In this current study, we performed ^{18}F-FDG PET/CT scans very strictly, this might be more difficult in multi-center setting and could result in an accumulation of small errors which could affect test-retest performance. Additionally differences in VOI definition of SUV_{peak} in our study versus those published elsewhere could partly explain reported differences in repeatability (*17*). In contrast to our study, in which we positioned a 1.2 cm^3 spheric VOI within tumor borders defined by a 50% isocontour to obtain the highest peak value, Weber et al. (*9*) placed a 1.5 cm diameter cylindric VOI in three consecutive axial slices over the voxel with the maximum uptake and the report did not specify whether a tumor border was defined. Therefore, there is a risk that nonmalignant tissue is taken into account when the maximum voxel is located near the edge of the tumor, and repeatability could be seriously affected by the variable location

of the maximum voxel, which is susceptible to noise (*18,19*). Furthermore, they assessed differences between test and retest scans after averaging all lesions in individual patients, because this may have a better correlation with patient outcome (*9*). Contrary to our results, no improvement in repeatability was found. Yet, improvement of variability would be expected if no systematic difference between both baseline scans exists, since differences would be reduced by averaging the data.

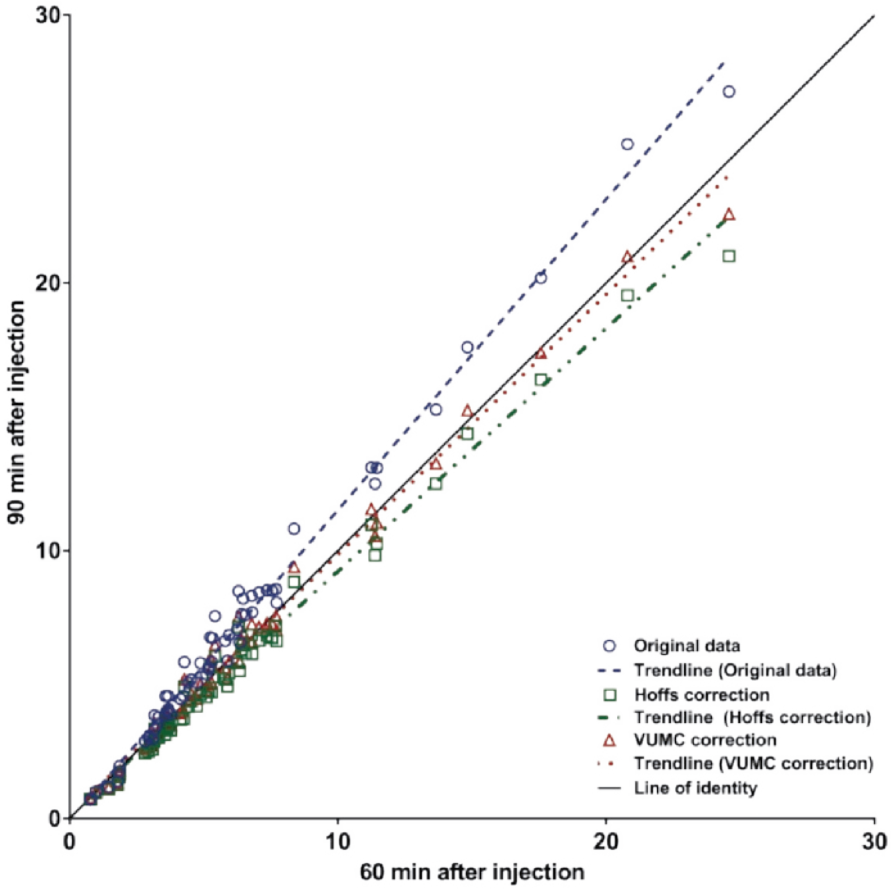


Figure 4.4: Different uptake time corrections applied on SUV_{peak} values of the 90 minutes data plotted against the 60 minutes data of the test scan.

Change in MATV has been shown to predict pathologic response in breast cancer after 2 cycles of chemotherapy, but few studies have assessed the repeatability of this

parameter (*20-22*). The repeatability of MATV was better in our study than in studies published by Frings et al. (*22*) in NSCLC patients and Hatt et al. (*20*) in esophageal cancer patients. These discrepancies are most likely explained by differences in uptake time (45 vs. 60 minutes p.i.) and VOI definition respectively. TLG has properties similar to MATV and showed similar repeatability, yet was influenced by uptake time. Two other studies investigated TLG repeatability in liver metastases and found RCs of 31.2% (*23,24*). The differences may partly be explained by differences in tumor type, as liver metastases tend to be more irregular compared to lung lesions and repeatability could be affected by higher background activity of the liver. Moreover, VOIs were delineated using a 41% (*24*) and 50% (*23*) isocontour corrected for local background based on SUV_{max} which could influence repeatability. Whether changes in FDG uptake metrics beyond the repeatability confidence intervals presented here also reflect sufficient clinical response remains to be shown.

Normalisation to Blood or Liver Uptake and Glucose Correction

Glucose correction deteriorated the test-retest performance in our study. Serum glucose levels were all within reference range and showed limited variability between the test and retest scan (< 2.2 mmol/L). Including this additional variable in the calculation of ^{18}F-FDG uptake metrics increases uncertainty and suggests that glucose correction should not be used when glucose levels are within reference range. Furthermore, we normalized SUV measures to liver and blood uptake to correct for inaccuracies in dose calibration, weight and length measurements and variations of tracer supply to the tumor (*3,25*). There was only little variability in the liver uptake between scans (1 ± 4%) yielding no improvement of repeatability. The same applies for the tumor-to-blood ratios in the 60 minutes data in contrast to expectations of van den Hoff et al. (*10*). Moreover, normalization to blood increased variability in the 90 minutes data, which might be explained by low count statistics (higher sensitivity to noise) of blood SUV at 90 minutes p.i..

Uptake Time Correction

Variation in uptake time has an important impact on the use of ^{18}F-FDG PET/CT as an imaging biomarker. With exception of MATV we have shown that uptake at 90 minutes p.i. is significantly higher compared to uptake at 60 minutes p.i. and therefore

supports the importance of timely procedures when quantitative measures are required. However, this can be difficult in clinical practice as shown in a study reporting a mean difference in uptake time of 33 ± 19 minutes (*26*). This underlines the need for methods to correct SUV for uptake time. In our study, we have chosen to prospectively evaluate the methods presented by van den Hoff et al. because they are based on known ^{18}F-FDG kinetics and therefore fundamentally attractive (*14*). Use of these methods improved correlation between the 60 and 90 minutes data, but adjustment of the *b*-value was required to offset underestimation of the 90 min corrected SUV. This suggests that the original method presented for uptake time correction of SUV may not be directly applicable and requires (further) validation. Moreover, it would be interesting to assess the effects of using this method to the multi-center data presented by Weber *et al.* to see if multi-center repeatability improves when correcting for uptake time variations (*9*). The higher *b*-value found in our dataset implies that the arterial input function decreases at a faster rate compared to the data presented by van den Hoff et al. (*14*). This discrepancy might be explained by differences in patient preparation (e.g. longer fasting prior to the scan or better prehydration and therefore more excretion). Moreover, in our study patients were scanned at two separate days and could therefore be affected by physiologic differences, but this has to be further explored. Despite the need for (minor) adjustment of one of the parameters of the uptake time correction method, we found a good correspondence between 90 and 60 min p.i. uptake metrics without affecting repeatability and therefore we propose further evaluation as a potential strategy to compensate for unwanted variability in uptake times during longitudinal studies.

Limitations

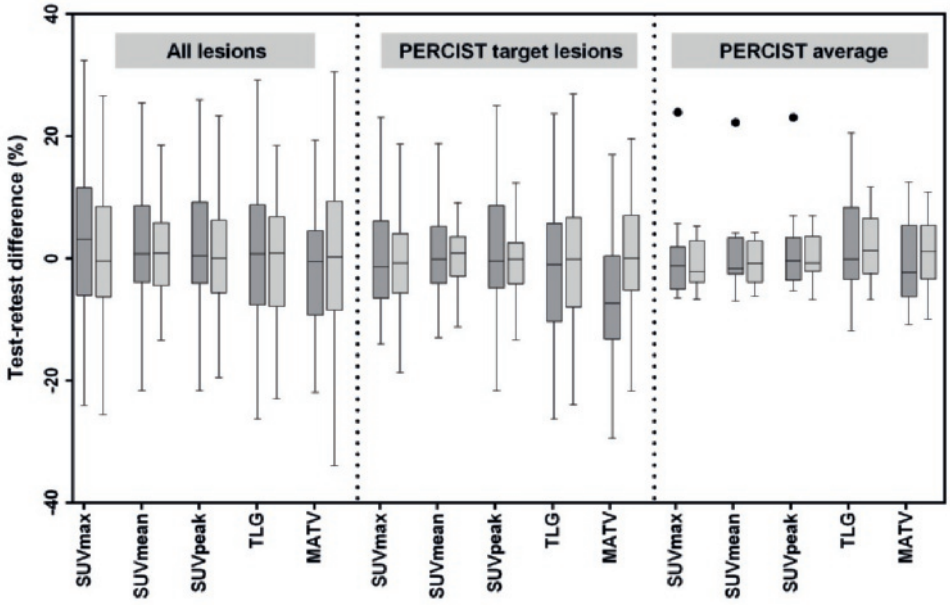
The main limitation to this study is the relatively small sample size. Unfortunately patient burden due to long scan time (± 60 minutes), because of the 60 and 90 minutes acquisitions, limited the collection of large datasets. However, to our knowledge no other studies have assessed these issues in such a comprehensive study design. Ideally, large multicentre trials should confirm our results, but implementation of this protocol into a trial would significantly increase the patient burden and would be less feasible compared with a repeatability study at one time-point only.

Second, we only assessed NSCLC patients, possibly limiting extrapolation to other tumor types. However, also extrathoracic lesions were included and there were no differences in repeatability depending on tumor location.

CONCLUSIONS

The results of this prospective study suggest that if up to 5 PERCIST target lesions are included, a 15% change of SUV_{mean} or SUV_{peak} reflects true metabolic response in patients with advanced NSCLC. If response is assessed using the averaged PERCIST target lesions this threshold could even be set at a < 10% change. No differences in test-retest performances were observed at 60 and 90 minutes post-injection and normalization to blood or liver uptake did not improve repeatability. Whether the thresholds found in this study are also valid in well-controlled multicenter studies, remains to be shown.

SUPPLEMENTAL MATERIALS



Supplemental Figure 4.1: Boxplots of the percentage differences between the test and retest scan as obtained from the 60 (dark) and 90 (light) minutes data with the outlier included in the analyses. The effect of lesion selection and averaging is shown on the different ^{18}F-FDG uptake metrics.

Supplemental Table 4.1: The mean relative differences and repeatability coefficients (RC) for several uptake metrics with the influence of different uptake times and lesions selection with the outlier included.

		All lesions		*PERCIST-criteria **		*PERCIST-criteria averaged*	
Uptake time	*Quantitative tracer uptake measures*	*Mean diffference (%)*	*RC (%)*	*Mean diffference (%)*	*RC (%)*	*Mean diffference (%)*	*RC (%)*
60 minutes post-injection	SUV_{max}	4.6	28.9	1.5	21.7	1.0	17.2
	SUV_{mean}	3.1	21.9	1.4	20.5	1.0	15.9
	SUV_{peak}	3.1	23.3	1.5	22.1	1.7	16.4
	TLG	0.6	30.8	0.8	22.7	1.9	18.9
	MATV	-2.4	30.0	-0.6	21.1	-0.5	14.2
90 minutes post-injection	SUV_{max}	0.0	22.8	-1.0	16.0	-1.0	8.0
	SUV_{mean}	0.2	17.5	-0.5	12.1	-0.8	7.1
	SUV_{peak}	0.0	16.1	0.0	12.4	0.0	7.8
	TLG	0.0	24.5	1.4	18.3	2.0	11.3
	MATV	-0.1	30.8	2.0	22.5	1.0	13.2

** PERCIST-criteria: 5 hottest lesions, max. 2 per organ and minimum SUV >1.5 x mean liver SUV + 2SDs*

REFERENCES

1. Truong MT, Viswanathan C, Erasmus JJ. Positron emission tomography/computed tomography in lung cancer staging, prognosis, and assessment of therapeutic response. *J Thorac Imaging*. 2011;26:132-146.
2. Boellaard R. Standards for PET image acquisition and quantitative data analysis. *J Nucl Med*. 2009;50(suppl 1):11S-20S.
3. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *J Nucl Med*. 2009;50(suppl 1):122S-150S.
4. van HM, Omloo JM, van Berge Henegouwen MI, et al. Fluorodeoxyglucose positron emission tomography for evaluating early response during neoadjuvant chemoradiotherapy in patients with potentially curable esophageal cancer. *Ann Surg*. 2011;253:56-63.
5. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45:228-247.
6. Barrington SF, Mikhaeel NG, Kostakoglu L, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group. *J Clin Oncol*. 2014;32:3048-3058.
7. de Langen AJ, Vincent A, Velasquez LM, et al. Repeatability of 18F-FDG uptake measurements in tumors: a metaanalysis. *J Nucl Med*. 2012;53:701-708.
8. Rockall AG, Avril N, Lam R, et al. Repeatability of quantitative FDG-PET/CT and contrast-enhanced CT in recurrent ovarian carcinoma: test-retest measurements for tumor FDG uptake, diameter, and volume. *Clin Cancer Res*. 2014;20:2751-2760.
9. Weber WA, Gatsonis CA, Mozley PD, et al. Repeatability of 18F-FDG PET/CT in advanced non-small cell lung cancer: prospective assessment in two multicenter trials. *J Nucl Med*. 2015;56:1137-1143.
10. van den Hoff J, Hofheinz F. Letter: Repeatability of tumor SUV quantification: the role of variable blood SUV. *J Nucl Med*. 2015;56:1635-1636.
11. Weber WA, Gatsonis C, Siegel B. Reply: Repeatability of tumor SUV quantification: the role of variable blood SUV. *J Nucl Med*. 2015;56:1636.
12. Frings V, van Velden FH, Velasquez LM, et al. Repeatability of metabolically active tumor volume measurements with FDG PET/CT in advanced gastrointestinal malignancies: a multicenter study. *Radiology*. 2014;273:539-548.
13. Boellaard R, Delgado-Bolton R, Oyen WJ, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. *Eur J Nucl Med Mol Imaging*. 2015;42:328-354.

14. van den Hoff J, Lougovski A, Schramm G, et al. Correction of scan time dependence of standard uptake values in oncological PET. *EJNMMI Res*. 2014;4:18.
15. Young H, Baum R, Cremerius U, et al. Measurement of clinical and subclinical tumour response using [18F]-fluorodeoxyglucose and positron emission tomography: review and 1999 EORTC recommendations. European Organization for Research and Treatment of Cancer (EORTC) PET Study Group. *Eur J Cancer*. 1999;35:1773-1782.
16. Velasquez LM, Boellaard R, Kollia G, et al. Repeatability of 18F-FDG PET in a multicenter phase I study of patients with advanced gastrointestinal malignancies. *J Nucl Med*. 2009;50:1646-1654.
17. Vanderhoek M, Perlman SB, Jeraj R. Impact of the definition of peak standardized uptake value on quantification of treatment response. *J Nucl Med*. 2012;53:4-11.
18. Lodge MA, Chaudhry MA, Wahl RL. Noise considerations for PET quantification using maximum and peak standardized uptake value. *J Nucl Med*. 2012;53:1041-1047.
19. Boellaard R, Krak NC, Hoekstra OS, Lammertsma AA. Effects of noise, image resolution, and ROI definition on the accuracy of standard uptake values: a simulation study. *J Nucl Med*. 2004;45:1519-1527.
20. Hatt M, Groheux D, Martineau A, et al. Comparison between 18F-FDG PET image-derived indices for early prediction of response to neoadjuvant chemotherapy in breast cancer. *J Nucl Med*. 2013;54:341-349.
21. Soussan M, Cyrta J, Pouliquen C, et al. Fluorine 18 fluorodeoxyglucose PET/CT volume-based indices in locally advanced non-small cell lung cancer: prediction of residual viable tumor after induction chemotherapy. *Radiology*. 2014;272:875-884.
22. Frings V, de Langen AJ, Smit EF, et al. Repeatability of metabolically active volume measurements with 18F-FDG and 18F-FLT PET in non-small cell lung cancer. *J Nucl Med*. 2010;51:1870-1877.
23. van Velden FH, Nissen IA, Jongsma F, et al. Test-retest variability of various quantitative measures to characterize tracer uptake and/or tracer uptake heterogeneity in metastasized liver for patients with colorectal carcinoma. *Mol Imaging Biol*. 2014;16:13-18.
24 Heijmen L, de Geus-Oei LF, de Wilt JH, et al. Reproducibility of functional volume and activity concentration in 18F-FDG PET/CT of liver metastases in colorectal cancer. *Eur J Nucl Med Mol Imaging*. 2012;39:1858-1867.
25. van den Hoff J, Oehme L, Schramm G, et al. The PET-derived tumor-to-blood standard uptake ratio (SUR) is superior to tumor SUV as a surrogate parameter of the metabolic rate of FDG. *EJNMMI Res*. 2013;3:77.

26. Kumar V, Nath K, Berman CG, et al. Variance of SUVs for FDG-PET/CT is greater in clinical practice than under ideal study settings. *Clin Nucl Med*. 2013;38:175-182.